View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Elsevier - Publisher Connector

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Multivariate Analysis 97 (2006) 1660–1674

Journal of Multivariate Analysis

www.elsevier.com/locate/jmva

Robust estimation of Cronbach's alpha

A. Christmann[a], S. Van Aelst[b,*]

[a] *University of Dortmund, Fachbereich Statistik, 44421 Dortmund, Germany*

[b] *Department of Applied Mathematics and Computer Science, Ghent University (UGENT), Krijgslaan 281 S9, B-9000 Gent, Belgium*

Received 28 September 2004

Available online 7 July 2005

Abstract

Cronbach's alpha is a popular method to measure reliability, e.g. in quantifying the reliability of a score to summarize the information of several items in questionnaires. The alpha coefficient is known to be non-robust. We study the behavior of this coefficient in different settings to identify situations where Cronbach's alpha is extremely sensitive to violations of the classical model assumptions. Furthermore, we construct a robust version of Cronbach's alpha which is insensitive to a small proportion of data that belong to a different source. The idea is that the robust Cronbach's alpha reflects the reliability of the bulk of the data. For example, it should not be possible that some small amount of outliers makes a score look reliable if it is not.

© 2005 Elsevier Inc. All rights reserved.

AMS 2005 subject classification: 62F35; 62H12; 62H20

Keywords: Cronbach's alpha; MCD; M-estimator; Robustness; S-estimator

1. Introduction

We consider the problem of measuring the reliability for a set of items such as in a test. Consider a series of items $Y_j = T_j + \varepsilon_j$ for $j = 1, \ldots, p$, where T_j are the true unobservable item scores and ε_j are the associated errors which are assumed to be independent from the true item scores and distributed with zero mean. The observed overall score of the p items is

* Corresponding author. Fax: +3292644995.

E-mail addresses: Christmann@statistik.uni-dortmund.de (A. Christmann), Stefan.VanAelst@UGent.be (S. Van Aelst).

0047-259X/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2005.05.012

given by $Y = Y_1 + \cdots + Y_p$ and the overall true but unobservable score is $T = T_1 + \cdots + T_p$. Reliability or consistency r of the set of items is defined as the ratio of the variance of the true scores to the total observed variance, that is $r = \mathrm{Var}(T)/\mathrm{Var}(Y)$. Since $\mathrm{Var}(T)$ cannot be calculated directly, measures to estimate the reliability r have been developed.

Cronbach [4] proposed the coefficient alpha as a measure of reliability in classical test theory (see also [15]). Cronbach's alpha estimates the consistency between items in a test, that is the internal consistency of the test. It is defined as

$$\alpha^C = \frac{p}{p-1} \frac{\mathrm{Var}\left(\sum_{j=1}^p Y_j\right) - \sum_{j=1}^p \mathrm{Var}\left(Y_j\right)}{\mathrm{Var}\left(\sum_{j=1}^p Y_j\right)} = \frac{p}{p-1} \frac{\sum \sum_{j \neq k} \sigma_{jk}}{\sum \sum_{j,k} \sigma_{jk}}, \tag{1}$$

where σ_{jk} is the covariance of the pair (Y_j, Y_k). It has been shown in [10] that Cronbach's alpha is always a lower bound of the reliability r. The values 0.7 or 0.75 are often used as cutoff value for Cronbach's alpha and thus for the reliability of the test. Cronbach's alpha has been investigated further in, e.g. [9,32,14,2].

Cronbach's alpha can be estimated by substituting empirical variances and covariances in (1) above. However, it is well known that classical estimators such as empirical variances and covariances can be heavily influenced by a few erroneous observations (see e.g. [11]). Therefore, the resulting estimate of Cronbach's alpha can be completely misleading as soon as some mistaken observations are present. To avoid this problem we propose a robust Cronbach's alpha estimate that is able to resist outliers and thus measures the internal consistency of the most central part of the observations. A robust measure of reliability was already proposed by Wilcox [34] who used the midvariance and midcovariance as robust estimates for the variances and covariances in (1). In this paper we propose to estimate the covariance matrix of $(Y_1, \ldots, Y_p)^t$ using a robust estimator and then we substitute the elements of this robust covariance estimate in (1).

Many robust estimators of multivariate location and scatter have been investigated in the literature, such as M-estimators [19,12], the minimum volume ellipsoid and minimum covariance determinant estimator [24], and S-estimators [8,25,16]. Recently, robust multivariate statistical methods based on robust estimation of location and scatter have been developed and investigated such as factor analysis [20], principal component analysis [7,30], canonical correlation analysis [5,31] and multivariate regression [26,33,1]. See also [21] for an overview. An advantage of constructing a robust Cronbach's alpha as proposed in this paper is that it can be obtained immediately from the robust scatter matrix estimate computed for the robust multivariate analysis without any additional computational load. This is a clear advantage over the proposal of Wilcox [34] that has to be computed separately and does not take into account the multivariate nature of the data.

In Section 2 we review robust estimators of multivariate location and scatter. The robust Cronbach's alpha is introduced in Section 3 where we also investigate some important properties. Section 4 contains results of simulation studies which show that the robust Cronbach's alpha performs well both in situations without and with outliers. A real data example is given in Section 5 and Section 6 concludes.

2. Robust estimators of location and scatter

The robust Cronbach's alpha can be computed from any robust scatter estimate. In this paper we will mainly use the reweighted minimum covariance determinant (RMCD) estimator and S-estimators which are highly robust estimators that can be computed with standard statistical software packages such as S-PLUS or SAS.

Consider a multivariate data set $\{y_i; 1 \leqslant i \leqslant n\}$ with $y_i = (y_{i1}, \ldots, y_{ip})^t \in \mathbb{R}^p$. Fix $\lceil n/2 \rceil \leqslant h \leqslant n$, where $\lceil n/2 \rceil$ denotes the smallest integer greater than or equal to $n/2$. Then the MCD looks for the subset $\{y_{i_1}, \ldots, y_{i_h}\}$ of size h which is the most concentrated subset of size h in the sense that its covariance matrix has the smallest determinant. The estimate for the center, t_n^0, is then the mean of the optimal subset and the covariance estimate C_n^0 is a multiple of the empirical covariance estimate based on the data in the optimal subset.

The breakdown value of an estimator is the smallest fraction of observations that needs to be replaced by arbitrary values to make the estimator useless (i.e. its norm goes to infinity), see e.g. [25]. We will denote $\gamma = (n-h)/n$ so that $0 \leqslant \gamma \leqslant 0.5$. It then follows that the MCD has breakdown value equal to γ. This means that a fraction γ of the data points may contain errors without having an unbounded effect on the MCD estimates. Moreover, the MCD location and scatter estimators are asymptotically normal and have a bounded influence function [3,6] which means that a small amount of contamination at a certain place can only have a bounded effect on the MCD estimates, see [11] for more information on the influence function. Two common choices for the subset size h are $h = [(n + p + 1)/2] \approx n/2$ (so $\gamma \approx 0.5$) which yields the highest possible breakdown value, and $h \approx 3n/4$ (i.e. $\gamma \approx 0.25$) which gives a better compromise between efficiency and breakdown. We will mainly use the 25% breakdown MCD, that is $h \approx 3n/4$, in this paper.

To increase the performance of the MCD it is customary to compute the reweighted MCD estimates (t_n^1, S_n^1) which are defined as

$$t_n^1 = \frac{\sum_{i=1}^n w(d_i^2) y_i}{\sum_{i=1}^n w(d_i^2)} \quad \text{and} \quad C_n^1 = \frac{\sum_{i=1}^n w(d_i^2)(y_i - t_n^1)(y_i - t_n^1)^t}{\sum_{i=1}^n w(d_i^2)}.$$

The weights $w(d_i^2)$ are computed as $w(d_i^2) = I(d_i^2 \leqslant q_\delta)$ where $q_\delta = \chi^2_{p,1-\delta}$ and $d_i^2 = (y_i - t_n^0)^t (C_n^0)^{-1} (y_i - t_n^0)$ is the squared robust distance of observation y_i based on the initial MCD estimates (t_n^0, C_n^0). It is customary to take $\delta = 0.025$, see [28]. The reweighted MCD estimators (RMCD) preserve the breakdown value [18] and the bounded influence function [17] of the initial MCD estimators but have a higher efficiency as shown in [6]. Recently, Rousseeuw and Van Driessen [27] constructed a fast algorithm to compute the RMCD.

S-estimates of location and scatter are defined as the couple (t_n^S, C_n^S) that minimizes $\det(C_n)$ under the constraint

$$\frac{1}{n} \sum_{i=1}^n \rho\left(\sqrt{(y_i - t_n)^t C_n^{-1} (y_i - t_n)}\right) \leqslant b,$$

over all $t_n \in \mathbb{R}^p$ and $C_n \in \mathrm{PDS}(p)$, where $\mathrm{PDS}(p)$ is the set of all positive definite symmetric matrices of size p. See e.g. [16] for important conditions on the ρ function. The

constant b satisfies $0 < b < \rho(\infty)$ and determines the breakdown value of the estimator which equals $\min(\frac{b}{\rho(\infty)}, 1-\frac{b}{\rho(\infty)})$ (see [16]). In this paper we usually select b such that the S-estimator has a 25% breakdown value. The most popular choice of ρ function is Tukey's biweight function which is given by

$$\rho_c(t) = \min\left(\frac{t^2}{2} - \frac{t^4}{2c^2} + \frac{t^6}{6c^4},\ \frac{c^2}{6}\right), \quad t \in \mathbb{R}. \tag{2}$$

Its derivative is given by

$$\psi_c(t) = t\left(1 - \frac{t^2}{c^2}\right)^2 I(|t| < c), \quad t \in \mathbb{R},$$

where I denotes the indicator function. The tuning constant c in the ρ function (2) can be selected such that consistency at a specific model distribution is obtained. S-estimators are asymptotically normal and have a bounded influence function [8,16]. Efficient algorithms to compute S-estimators have been constructed in [29,23]. The 25% breakdown S-estimator of the scatter matrix based on Tukey's biweight function will be denoted S_{bw}.

Another class of robust scatter matrix estimators are M-estimators. We will consider the M-estimator based on the assumption of Student's t_3 distribution which will be denoted by T3. It is obtained as the solution of the estimating equations

$$t_n^{\mathrm{T3}} = \left(\sum_{i=1}^n w_i y_i\right) \Big/ \left(\sum_{i=1}^n w_i\right) \text{ and } C_n^{\mathrm{T3}} = \frac{1}{n}\sum_{i=1}^n w_i (y_i - t_n^{\mathrm{T3}})(y_i - t_n^{\mathrm{T3}})^t,$$

where $w_i = (3+p)/(3+d_i^2)$ with $d_i^2 = (y_i - t_n^{\mathrm{T3}})^t (C_n^{\mathrm{T3}})^{-1}(y_i - t_n^{\mathrm{T3}})$. The T3 estimator not only has reasonable robustness and efficiency properties, but also some additional advantages. There exists a unique solution of the objective criterion under very weak assumptions and there also exists an always converging iterative algorithm to compute the estimate, as was shown in [12,13]. Furthermore, this estimator is intuitively appealing as it is a maximum likelihood estimator if the errors follow a multivariate t_3 distribution. However, the main disadvantage of T3 is its low breakdown point.

3. Robust Cronbach's alpha

Consider a data set $Y_n = \{y_i; i = 1, \ldots, n\} \subset \mathbb{R}^p$ and denote by C_n the corresponding scatter estimate such as the empirical covariance S, RMCD, S_{bw} or T3 estimate of scatter. Then the corresponding Cronbach's alpha estimate is defined as

$$\alpha_C(Y_n) = \frac{p}{p-1}\,\frac{\sum\sum_{j\neq k} c_{jk}}{\sum\sum_{j,k} c_{jk}}, \tag{3}$$

where c_{ij}, $i, j = 1, \ldots, p$, are the elements of the matrix C_n and C indicates S, MCD, RMCD, S_{bw} or T3. When using the empirical covariance matrix S in (3) we obtain the

classical estimate of Cronbach's alpha derived from (1). On the other hand, using a robust estimate of the covariance matrix in (3) will lead to a robust estimate of Cronbach's alpha.

Let the observed item scores $(Y_1, \ldots, Y_p)$ have a distribution $F_{\mu,\Sigma}$ which belongs to the class of unimodal elliptically symmetric distributions. Hence, the density function is of the form

$$f_{\mu,\Sigma}(y) = \frac{g((y-\mu)^t \Sigma^{-1}(y-\mu))}{\sqrt{\det(\Sigma)}}$$

with $\mu \in \mathbb{R}^p$ and $\Sigma \in \mathrm{PDS}(p)$ and where the function g has a strictly negative derivative. Multivariate normal distributions obviously belong to this class of distributions. With $\Sigma = (\sigma_{ij})$, we then focus on estimating the quantity

$$\alpha = \frac{p}{p-1} \frac{\sum \sum_{j \neq k} \sigma_{jk}}{\sum \sum_{j,k} \sigma_{jk}}.$$

If the scatter estimator C_n is consistent in probability or almost surely, then it follows immediately from Slutsky's theorem that the corresponding Cronbach's alpha estimator given by (3) is a consistent estimator of α (in probability or almost surely). Consistency of robust location/scatter estimators at elliptically symmetric distributions has been shown in [3] for the MCD, in [17] for the RMCD and in [8,16] for S-estimators.

The influence function (IF) describes the local robustness of the functional version of an estimator. A statistical functional corresponding to an estimator C_n is a map C which maps any p-variate distribution G on $C(G) \in \mathrm{PDS}(p)$ such that $C(F_n) = C_n$ for any empirical distribution function F_n. The functional version of Cronbach's alpha associated with a scatter functional $C(G)$ will be denoted by $\alpha_C(G)$. It follows immediately that $\alpha_C(F_{\mu,\Sigma}) = \alpha$ whenever $C(F_{\mu,\Sigma}) = \Sigma$, that is, whenever C is Fisher-consistent at elliptical distributions $F_{\mu,\Sigma}$. The MCD and RMCD scatter estimators can be made Fisher-consistent at elliptical distributions by using a suitable multiplication factor in the definition of C_n^0 and C_n^1 (see e.g. [6,22]). Similarly, the tuning constant c in the ρ function (2) can be selected such that S_{bw} is Fisher-consistent at a specific elliptical model distribution.

The influence function of the functional α_C at the distribution $F_{\mu,\Sigma}$ measures the effect on $\alpha_C(F_{\mu,\Sigma})$ of adding a small mass at a certain point y. Such a perturbation mimics the occurrence of isolated outliers, e.g. due to typing errors. Hence, a robust method should have a bounded influence function such that contamination at any point can only have a limited effect on the estimate. If we denote by Δ_y the distribution putting all its mass on y, then the influence function is given by

$$\begin{aligned} \mathrm{IF}(y; \alpha_C, F_{\mu,\Sigma}) &= \lim_{\varepsilon \downarrow 0} \frac{\alpha_C((1-\varepsilon)F_{\mu,\Sigma} + \varepsilon\Delta_y) - \alpha_C(F_{\mu,\Sigma})}{\varepsilon} \\ &= \frac{\partial}{\partial \varepsilon} \alpha_C((1-\varepsilon)F_{\mu,\Sigma} + \varepsilon\Delta_y)|_{\varepsilon=0}. \end{aligned} \tag{4}$$

See [11] for further details. For scatter matrix estimators $C(G)$ that are Fisher-consistent at elliptically symmetric distributions $F := F_{\mu,\Sigma}$ and possess an influence function, combining

the functional version of (3) with (4) yields

$$\begin{aligned}
\mathrm{IF}(y;\alpha_C,F) &= \frac{p}{p-1}\,\frac{\partial}{\partial\varepsilon}\,\frac{\sum\sum_{j\neq k} c_{jk}(F_\varepsilon)}{\sum\sum_{j,k} c_{jk}(F_\varepsilon)}\Big|_{\varepsilon=0} \\
&= \frac{p}{p-1}\,\frac{\sum\sum_{j\neq k}\mathrm{IF}(y;c_{jk},F)}{\sum\sum_{j,k} c_{jk}(F)} \\
&\quad - \frac{p}{p-1}\,\frac{\left(\sum\sum_{j\neq k} c_{jk}(F)\right)\left(\sum\sum_{j,k}\mathrm{IF}(y;c_{jk},F)\right)}{\left(\sum\sum_{j,k} c_{jk}(F)\right)^2} \\
&= \frac{p}{p-1}\,\frac{\sum\sum_{j\neq k}\mathrm{IF}(y;c_{jk},F)}{\sum\sum_{j,k}\sigma_{jk}} - \alpha\,\frac{\sum\sum_{j,k}\mathrm{IF}(y;c_{jk},F)}{\sum\sum_{j,k}\sigma_{jk}}.
\end{aligned}$$

Hence, we obtain the following result.

Theorem 3.1. *If the scatter matrix estimator C possesses an influence function then the influence function of α_C at elliptically symmetric distributions $F := F_{\mu,\Sigma}$ is given by*

$$\mathrm{IF}(y;\alpha_C,F) = \frac{\frac{p}{p-1}\sum\sum_{j\neq k}\mathrm{IF}(y;c_{jk},F) - \alpha\sum\sum_{j,k}\mathrm{IF}(y;c_{jk},F)}{\sum\sum_{j,k}\sigma_{jk}}.$$

It follows that the influence function of Cronbach's alpha is bounded as soon as the influence function of the scatter matrix estimator is bounded which is the case for RMCD, T3, and S-estimators with bounded ρ function such as S_{bw}. Therefore, our approach based on a robust estimate of the scatter matrix indeed yields a robust estimate of Cronbach's alpha.

As an example, let us consider the influence function of the S-estimator of scatter based on Tukey's biweight function (2) for a multivariate standard normal distribution $F = N(\mathbf{0},\mathbf{I})$ which is given by

$$\mathrm{IF}(y;C^S,F) = \frac{2}{\gamma_3}\,(\rho(\|y\|)-b_0) + \frac{1}{\gamma_1}\,p\,\psi(\|y\|)\,\|y\|\left(\frac{yy^t}{\|y\|^2}-\frac{1}{p}\,\mathbf{I}\right),$$

where $\gamma_1 = (p+2)^{-1}\mathrm{E}_F\left[\psi'(\|Y\|)\,\|Y\|^2 + (p+1)\psi(\|Y\|)\,\|Y\|\right]$ and $\gamma_3 = \mathrm{E}_F\left[\psi(\|Y\|)\|Y\|\right]$ (see [16, Corollary 5.2]). The influence function of Cronbach's alpha based on $\mathrm{S_{bw}}$ for the bivariate standard normal distribution is given in Fig. 1. Note that the influence function is smooth and bounded. Furthermore, for points with large euclidean norm $\|y\|$ it is constant, but not necessarily equal to zero for general multivariate normal distributions. Hence, data points lying far away from the bulk of the data cloud only have a small impact on this robust Cronbach's alpha.

As the influence function is an asymptotical concept, it is also interesting to consider an empirical version of the influence function for finite sample sizes. Here, we consider the *sensitivity curve* SC_n, cf. [11, p. 93]. The sensitivity curve of Cronbach's alpha $\alpha_C(Y_n)$ given a multivariate data set $Y_n = (y_1,\ldots,y_n)$ is defined by

$$\mathrm{SC}_n(y) = n\,[\alpha_C(y_1,\ldots,y_n,y) - \alpha_C(y_1,\ldots,y_n)], \quad y\in\mathbb{R}^p.$$

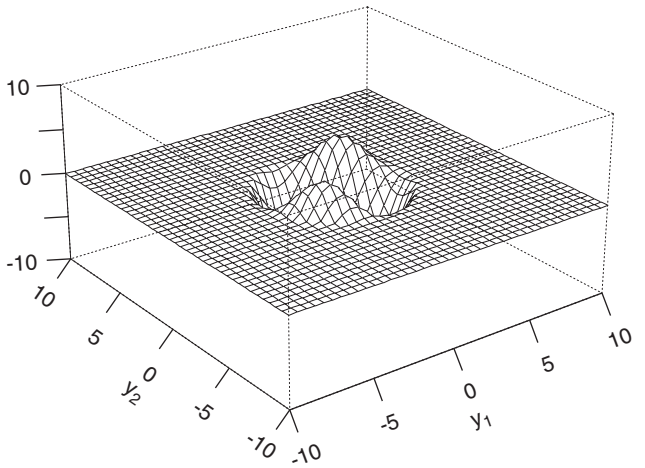

Fig. 1. Influence function of Cronbach's alpha based on the S-estimator S_{bw} at the bivariate normal distribution.

Hence, SC_n describes the standardized behavior of the estimate if one arbitrary data point y is added to the data set.

Sensitivity curves of Cronbach's alpha based on empirical (co)variances and its robust alternatives are given in Fig. 2 for the bivariate standard normal distribution. Note that due to different magnitudes of the sensitivity curves the scaling of the vertical axis in the plots is not identical for all four estimates. In Fig. 2, we consider Cronbach's alpha based on S, RMCD, S_{bw}, and T3. We see that the impact of even one single additional observation can be extremely large for the classical Cronbach's alpha based on S, whereas the robustifications behave much more stable. Especially the sensitivity curves based on RMCD and S_{bw} are very stable for observations far away from the bulk of the data. Note that the sensitivity curve of Cronbach's alpha based on S_{bw} is very similar to the influence function shown in Fig. 1, although we used only a moderate sample size of $n = 100$ to construct SC_n. Cronbach's alpha based on T3 shows a smooth and more robust behavior than the classical estimator, but it is not as robust as the estimators based on RMCD and S_{bw} for extreme outliers.

4. Simulations

We investigated the behavior of the classical and robust Cronbach's alpha estimators for finite samples via simulations for sample sizes of $n = 40$, 100, and 500. Let $(Y_1, \ldots, Y_p)^t$ be a random vector with multivariate distribution F. Since $Y_j = T_j + \varepsilon_j$ we have that $E(Y_j) = E(T_j) = \mu_j$, the expected value for item j. For dimension $p = 2$ we define location vectors $\mu = (0, 0)'$, $\mu_1 = (2, 2)'$, and $\mu_2 = (-2, 2)'$. For dimension $p = 10$ we define location vectors $\mu = \mathbf{0} \in \mathbb{R}^p$, $\mu_1 = (2, \ldots, 2)'$, and $\mu_2 = (-2, 2, \ldots, 2)'$. As scatter matrices we use $\Sigma = (\sigma_{ij}) \in \mathbb{R}^{p\times p}$, where $\sigma_{ij} = 1$, if $i = j$, and $\sigma_{ij} = \rho$, if $i \neq j$, and $\Sigma_1 = (\sigma_{ij}) \in \mathbb{R}^{p\times p}$, where $\sigma_{ij} = 1$, if $i = j$. If $p = 2$ the off-diagonal elements

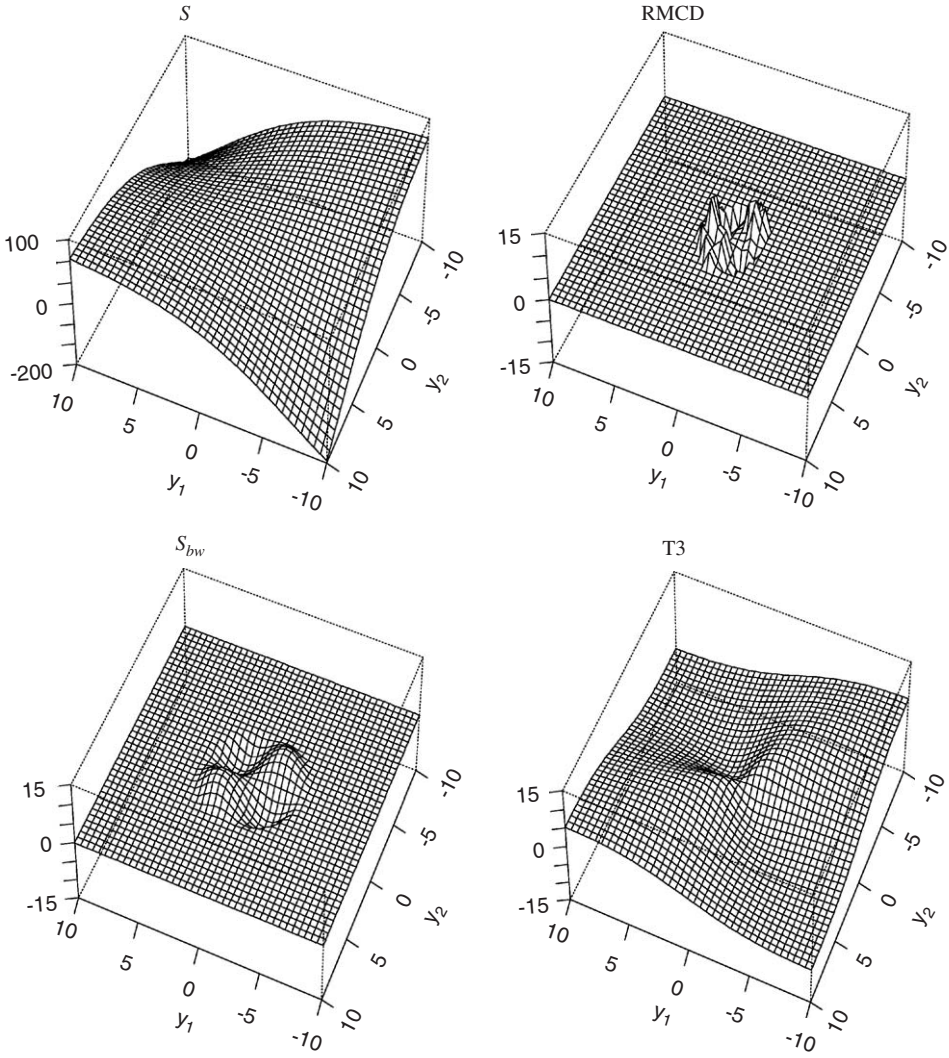

Fig. 2. Sensitivity curves for a two-dimensional data set with $n = 100$ observations simulated from $F = \mathrm{N}(\mathbf{0}, \mathbf{I})$.

of Σ_1 are $\sigma_{12} = \sigma_{21} = -\rho$. If $p = 10$ we set the off-diagonal elements of Σ_1 equal to $\sigma_{ij} = -\rho$, if $\{i = 1 \text{ or } j = 1 \text{ and } i \neq j\}$, and $\sigma_{ij} = \rho$, if $\{i > 1, j > 1 \text{ and } i \neq j\}$. We use $\delta = 0.05, 0.10,$ and 0.20 as contamination proportions, and study correlations of $\rho = 0, 0.2, 0.5,$ and 0.8. In the simulations the following five probability models are considered:

- N: multivariate normal: $F = N(\mu, \Sigma)$,
- t_3: multivariate Student's t with 3 df: $F = t_3(\mu, \Sigma)$,

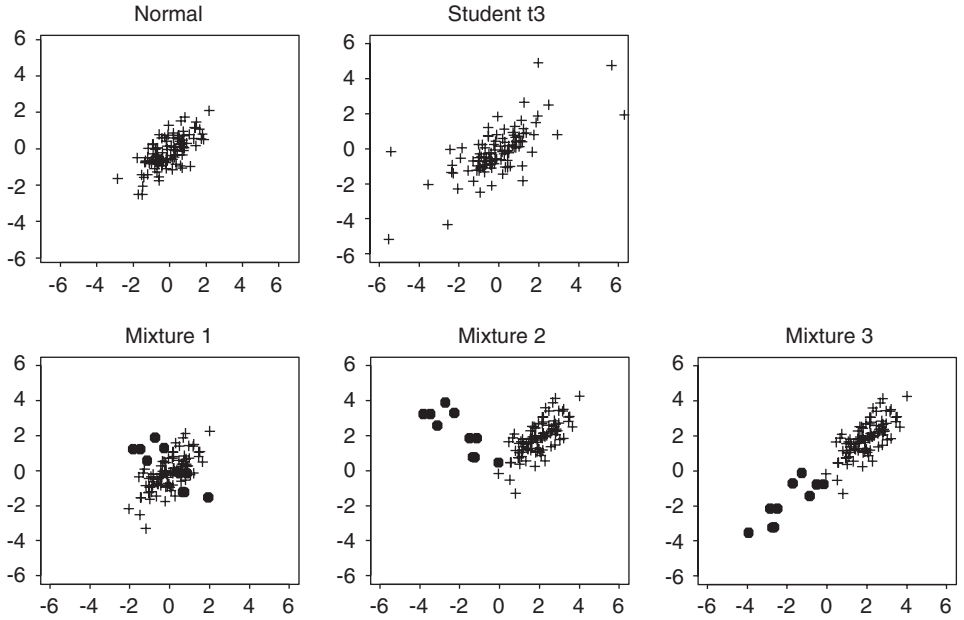

Fig. 3. Scatterplots of simulated data for $p = 2$, $n = 100$, $\rho = 0.8$, and $\delta = 10\%$.

- $\delta\%$ M1: contamination model 1 with different covariance matrix: $F = (1-\delta)N(\mu, \Sigma) + \delta N(\mu, \Sigma_1)$,
- $\delta\%$ M2: contamination model 2 with different location parameter and covariance matrix: $F = (1-\delta)N(\mu_1, \Sigma) + \delta N(\mu_2, \Sigma_1)$,
- $\delta\%$ M3: contamination model 3 with different location parameter: $F = (1-\delta)N(\mu_1, \Sigma) + \delta N(-\mu_1, \Sigma)$.

To allow a visual comparison of these probability models, scatterplots of data sets simulated according to these five models for $p = 2$, $n = 100$, $\rho = 0.8$, and $\delta = 10\%$ are given in Fig. 3. The contaminated data points are marked as dots. In the context of a questionnaire the contamination models can be explained as follows. Suppose a positive answer expresses to what extend the respondent '*agrees*' with the statement in an item and a negative answer indicates the amount of '*disagreement*'. The contamination in outlier model 1 can be caused by respondents that incorrectly reversed the statement of one item and hence give an answer that does not match with their answers to the other items. Contamination model 2 is the same but now the population average is not zero. Contamination model 3 expresses that some respondents reversed the scale of their answers in the whole questionnaire, that is they give negative answers when agreeing with the statement and vice versa.

For each simulation we generated 1000 data sets and computed bias and mean squared error of Cronbach's alpha based on the empirical covariance S and based on the robust alternatives MCD, RMCD, S_{bw}, all with 25% breakdown point, and T3. Moreover, in the simulations we also included the robust Cronbach's alpha based on midvariance and

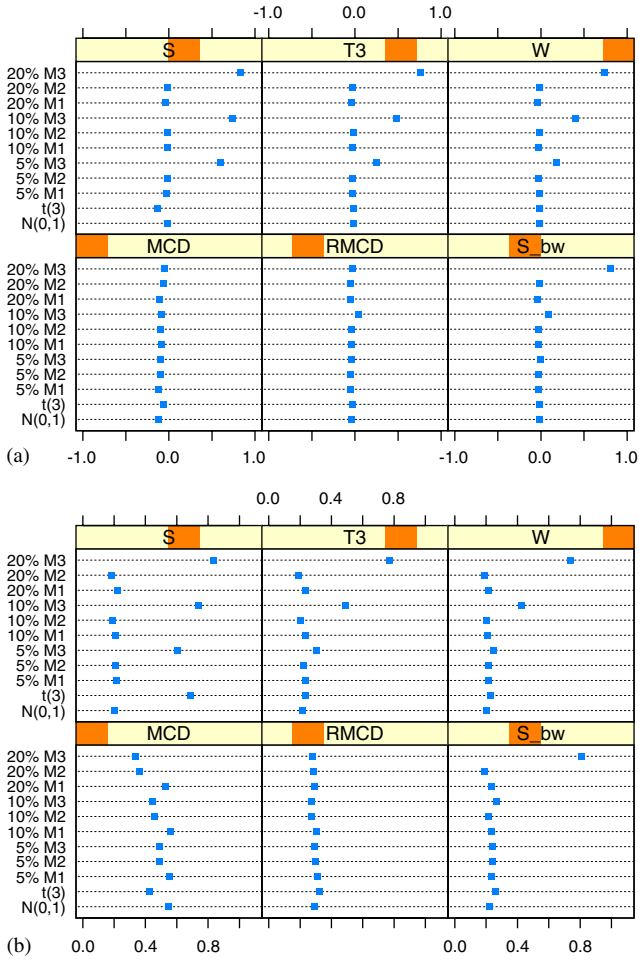

Fig. 4. (a) Bias and (b) square root of mean squared error for several estimators of Cronbach's α for $p = 2$, $\rho = 0$, and $n = 100$. Under classical normality assumptions the true value $\alpha = 0$.

midcovariance as proposed by Wilcox [34]. We denote this method by W. Some results of the simulations are summarized in Figs. 4 and 5 for $p = 2$ dimensions and in Fig. 6 for $p = 10$. The simulation results for the other situations were very similar.

First, note that these simulations confirm that the classical Cronbach's alpha is non-robust with respect to violations of the model assumptions. It can seriously overestimate (contamination model 3, Fig. 4a) or underestimate (contamination models 1 and 2, Fig. 5a) the value of the parameter α of the population. Student's distribution t_3 is elliptically symmetric with heavier tails than the normal distribution and is often a good approximation to the distribution of high quality data, cf. [11, p. 23]. However, even in this situation the bias and the MSE are often much larger than under the classical assumption (see e.g.

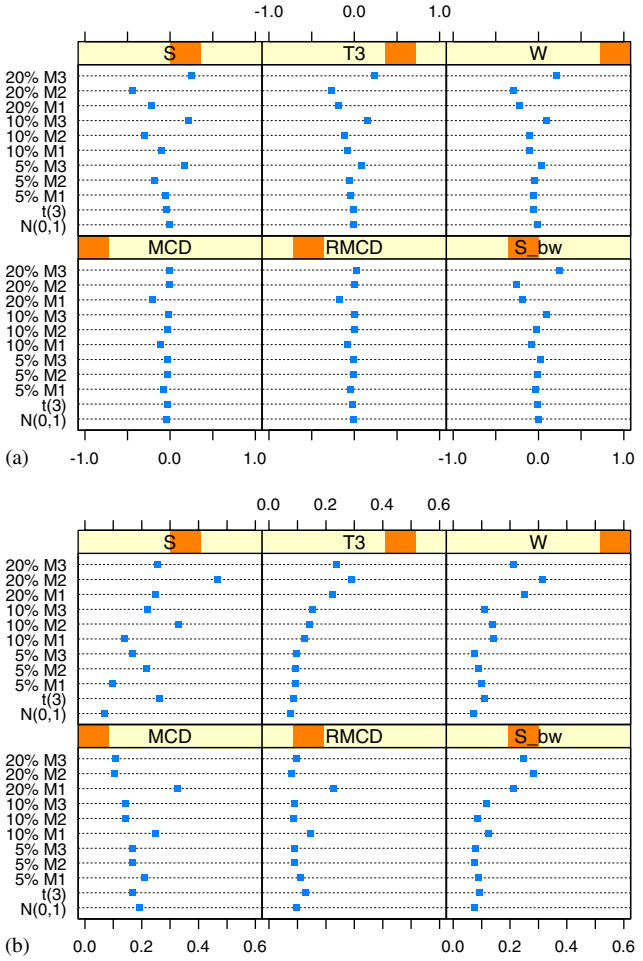

Fig. 5. (a) Bias and (b) square root of mean squared error for several estimators of Cronbach's α for $p = 2$, $\rho = 0.5$, and $n = 100$. Under classical normality assumptions the true value $\alpha = 0.667$.

Fig. 4). The same is true for contamination model 1 as can be seen in Fig. 5. If the contamination is asymmetric as in the other two contamination models, the behavior of Cronbach's alpha can be even worse.

Robust Cronbach's alpha based on all three robust covariance estimators yield more stable estimates than the classical approach. In most cases Cronbach's alpha based on RMCD gives a better result than the Cronbach's alpha based on the initial MCD estimator, which often has a higher bias and a higher mean squared error (see Figs. 4 and 5). Hence, we do not consider the MCD results in Fig. 6 anymore. Cronbach's alpha based on RMCD is the only estimator under consideration which still gives reasonable results if the mixing proportion is as high as $\delta = 20\%$. Furthermore, this estimator often gives already better results for the multivariate t_3 distribution.

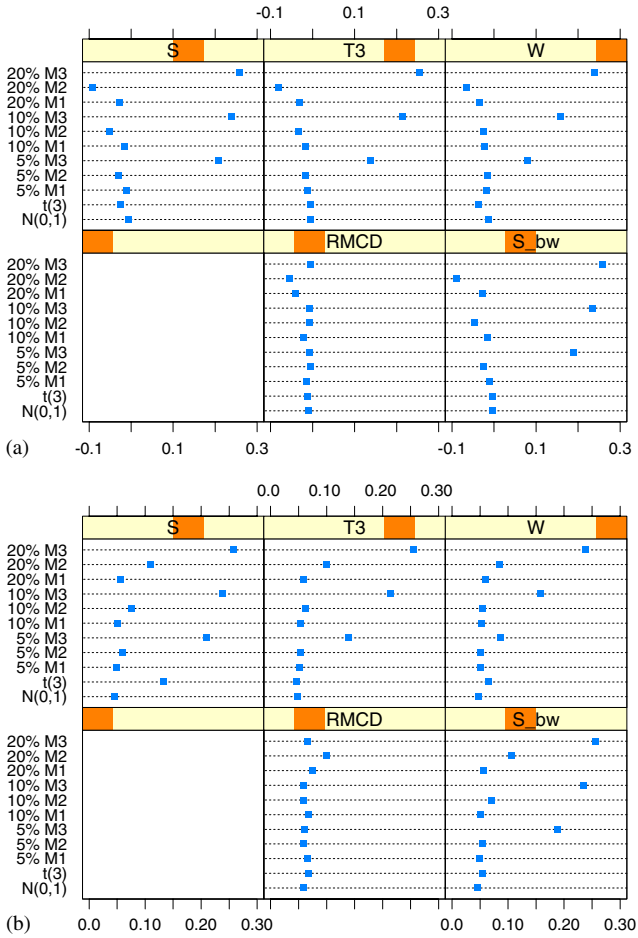

Fig. 6. (a) Bias and (b) square root of the mean squared error for $p = 10$, $\rho = 0.2$, and $n = 100$. Under classical normality assumptions the true value $\alpha = 0.714$.

When the assumption of normality is not valid, Cronbach's alpha based on the Tukey biweight S-estimator S_{bw}, performed best in many situations except for cases with contamination proportion $\delta = 20\%$. This amount of contamination is close to the breakdown point (25%) of the estimator and causes a large (but bounded) bias which seriously affects its performance. When the dimension increases, also the efficiency of the S-estimator increases, but the robustness decreases as can be seen from Fig. 6. This behavior has been noted before, see e.g. [6]. Finally, S_{bw} performs almost as good as the classical estimator if the assumption of normality is fulfilled.

The M-estimator T3 yields more robust results than the classical approach based on the empirical covariance matrix, but even for models with 5% of contamination it often gives worse results than the estimators based on RMCD or S_{bw}, especially for contamination

model 3 (see Fig. 5). This behavior of T3 coincides with the properties of the sensitivity curves shown in Section 3.

Finally, Wilcox' estimator usually behaves similar to the T3 estimator. Hence, although the midvariance has a high efficiency and bounded influence function, the resulting estimator often has a high bias when contamination is present in the data.

5. Example

To illustrate the usefulness of a robust Cronbach's alpha coefficient for a real data set, we investigate the internal consistency of scores obtained by 50 computer science students on 4 different projects in their third year of education at Ghent University.

The projects were scored on a scale of 0–20 and the grades could be specified up to the first decimal. Fig. 7 shows the robust distances of the observations based on RMCD with 50% breakdown point versus their index. The observations are ranked according to their overall average score. The horizontal line corresponds with the 97.5% percentile of the χ^2_4 distribution which is often used as a cutoff to detect outliers (see e.g. [28]). From this plot we can identify 8 outliers among the students with low overall average. The 50% breakdown point biweight S-estimator detected the same 8 outliers. To see how these outliers affect estimates of Cronbach's alpha we compare the estimates introduced before. Since we detected 16% of outliers in this dataset we use the 50% breakdown versions of RMCD and biweight S-estimators to avoid high bias due to the contamination. The Cronbach's alpha coefficients based on the empirical covariance S, Wilcox midvariance W and T3 were 0.72, 0.72 and 0.73, respectively. On the other hand, robust Cronbach's alpha coefficients based on RMCD and the biweight S-estimator both are 0.77. If 0.75 is used as a cutoff value for consistency, the outliers thus have a serious effect on the data analysis. To compare, we also computed the classical Cronbach's alpha for the data without the 8 outliers which yields 0.79. A closer examination of the data reveals that the outliers correspond to students that obtained at least one low score because they made only a small part of the project.

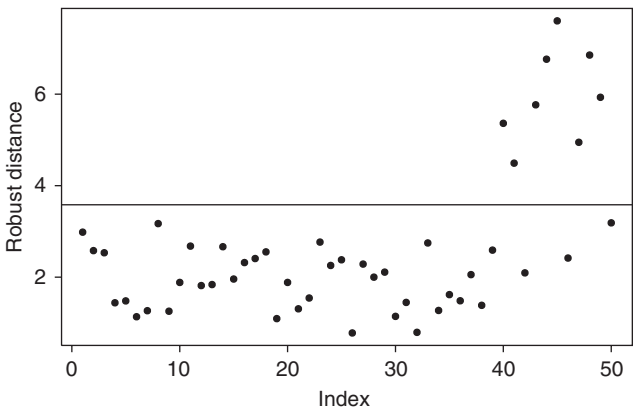

Fig. 7. Plot of robust distances based on RMCD versus the index.

6. Discussion

The reliability measure Cronbach's alpha is non-robust, even a single observation can have a high impact on this coefficient. Therefore, we proposed robust alternatives, which have good robustness properties, e.g. a bounded influence function, perform well in a simulation study with respect to bias and mean squared error, and are easy to compute with common statistical software packages such as SAS, S-PLUS or R. Software code to compute the robust Cronbach's alpha in SAS and S-PLUS is available from *http://www.statistik.uni-dortmund.de/sfb475/berichte/cronbach.zip* .

Acknowledgments

We thank Prof. David M. Rocke (University of California, Davis) for making available his program to compute the S-estimator. Andreas Christmann gratefully acknowledges the financial support of the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of complexity in multivariate data structures") and of the Forschungsband DoMuS (University of Dortmund). Stefan Van Aelst gratefully acknowledges the financial support of the Fund for Scientific Research—Flanders.

References

[1] J. Agulló, C. Croux, S. Van Aelst, The multivariate least trimmed squares estimator, (2002), submitted.
[2] G. Bravo, L. Potvin, Estimating the reliability of continuous measures with Cronbach's alpha or the intraclass correlation coefficient: toward the integration of two traditions, J. Clin. Epidemiol. 44 (1991) 381–390.
[3] R.W. Butler, P.L. Davies, M. Jhun, Asymptotics for the minimum covariance determinant estimator, Ann. Statist. 21 (1993) 1385–1400.
[4] L.J. Cronbach, Coefficient alpha and the internal structure of tests, Psychometrika 16 (1951) 297–334.
[5] C. Croux, C. Dehon, Analyse Canonique basèe sur des Estimateurs Robustes de la Matrice de Covariance, La Rev. Statist. Apl. 2 (2002) 5–26.
[6] C. Croux, G. Haesbroeck, Influence function and efficiency of the minimum covariance determinant scatter matrix estimator, J. Multivariate Anal. 71 (1999) 161–190.
[7] C. Croux, G. Haesbroeck, Principal component analysis based on robust estimators of the covariance or correlation matrix: influence function and efficiencies, Biometrika 87 (2000) 603–618.
[8] L. Davies, Asymptotic behavior of S-estimators of multivariate location parameters and dispersion matrices, Ann. Statist. 15 (1987) 1269–1292.
[9] L.S. Feldt, The approximate sampling distribution of Kuder–Richardson reliability coefficient twenty, Psychometrika 30 (1965) 357–370.
[10] L. Guttman, Reliability formulas that do not assume experimental independence, Psychometrika 18 (1953) 225–239.
[11] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, W.A. Stahel, Robust Statistics: The Approach based on Influence Functions, Wiley, New York, 1986.
[12] J.T. Kent, D.E. Tyler, Redescending M-estimates of multivariate location and scatter, Ann. Statist. 19 (1991) 2102–2119.
[13] J.T. Kent, D.E. Tyler, Y. Vardi, A curious likelihood identity for the multivariate T-distribution, Comm. Statist.—Simul. 23 (1994) 441–453.
[14] H.C. Kraemer, Extension of Feldt's approach to testing homogeneity of coefficients of reliability, Psychometrika 46 (1981) 41–45.

[15] G.F. Kuder, M.W. Richardson, The theory of the estimation of test reliability, Psychometrika 2 (1937) 151–160.
[16] H.P. Lopuhaä, On the relation between S-estimators and M-estimators of multivariate location and covariance, Ann. Statist. 17 (1989) 1662–1683.
[17] H.P. Lopuhaä, Asymptotics of reweighted estimators of multivariate location and scatter, Ann. Statist. 27 (1999) 1638–1665.
[18] H.P. Lopuhaä, P.J. Rousseeuw, Breakdown points of affine equivariant estimators of multivariate location and covariance matrices, Ann. Statist. 19 (1991) 229–248.
[19] R.A. Maronna, Robust M-estimates of multivariate location and scatter, Ann. Statist. 4 (1976) 51–67.
[20] G. Pison, P.J. Rousseeuw, P. Filzmoser, C. Croux, Robust factor analysis, J. Multivariate Anal. 84 (2003) 145–172.
[21] G. Pison, S. Van Aelst, Diagnostic plots for robust multivariate methods, J. Comput. Graphical Statist. 13 (2004) 310–329.
[22] G. Pison, S. Van Aelst, G. Willems, Small sample corrections for LTS and MCD, Metrika 55 (2002) 111–123.
[23] D.M. Rocke, D.L. Woodruff, Computation of robust estimates of multivariate location and shape, Statist. Neerlandica 47 (1993) 27–42.
[24] P.J. Rousseeuw, Least median of squares regression, J. Amer. Statist. Assoc. 79 (1984) 871–880.
[25] P.J. Rousseeuw, A.M. Leroy, Robust Regression and Outlier Detection, Wiley, New York, 1987.
[26] P.J. Rousseeuw, S. Van Aelst, K. Van Driessen, J. Agulló, Robust multivariate regression, Technometrics 46 (2004) 293–305.
[27] P.J. Rousseeuw, K. Van Driessen, A fast algorithm for the minimum covariance determinant estimator, Technometrics 41 (1999) 212–223.
[28] P.J. Rousseeuw, B.C. van Zomeren, Unmasking multivariate outliers and leverage points, J. Amer. Statist. Assoc. 85 (1990) 633–651.
[29] D. Ruppert, Computing S-estimators for regression and multivariate location/dispersion, J. Comput. Graphical Statist. 1 (1992) 253–270.
[30] M. Salibian-Barrera, S. Van Aelst, G. Willems, PCA based on multivariate MM-estimators with fast and robust bootstrap, (2005), submitted.
[31] S. Taskinen, C. Croux, A. Kankainen, E. Ollila, H. Oja, Influence functions and efficiencies of the canonical correlation and vector estimates based on scatter and shape matrices, J. Multivariate Anal. (2005), to appear.
[32] J.M.F. Ten berge, F.E. Zegers, A series of lower bounds to the reliability of a test, Psychometrika 43 (1978) 575–579.
[33] S. Van Aelst, G. Willems, Multivariate regression S-estimators for robust estimation and inference, Statist. Sinica, (2005), to appear.
[34] R.R. Wilcox, Robust generalizations of classical test reliability and Cronbach's alpha, British J. Math. Statist. Psychol. 45 (1992) 239–254.